ИНФОРМАТИКА

УДК 539.3

В. М. Буре[1,2], *О. А. Митрофанова*[1]

ПРОГНОЗ ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ ЭКОЛОГИЧЕСКИХ ДАННЫХ С ПРИМЕНЕНИЕМ КРИГИНГА И БИНАРНОЙ РЕГРЕССИИ

[1] Агрофизический научно-исследовательский институт, Российская Федерация, 195220, Санкт-Петербург, Гражданский пр., 14
[2] Санкт-Петербургский государственный университет, Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Существует ряд экологических задач, связанных с прогнозом пространственного распределения экологических параметров. В работе рассматривается одна из таких задач. Предполагается, что исходными данными являются набор экологических или агрохимических данных, измеренных контактным способом (например, показания N-тестера интенсивности окраски листьев растений), а также аэрофотоснимок обследуемого объекта (например, поля). Необходимо оценить пространственное распределение экологического параметра. В статье предложен подход к решению задачи с совместным использованием методов кригинга и бинарной регрессии. Предварительно с помощью метода классификации можно определить однородные зоны поля (кластеры) на снимке. Предполагается, что в каждой выделенной зоне имеется набор экологических данных. В дальнейшем изучается каждая зона отдельно. Необходимо оценить уровень показателя в рассматриваемой зоне. Вначале проводится вариограммный анализ, строится модель вариограммы. Далее строится набор оценок экологического параметра с помощью метода ординарного кригинга. После этого задается пороговое значение экологического параметра для рассматриваемой зоны, вводится фиктивная переменная, которая принимает значение 1, если величина параметра превысила пороговую, и 0 в ином случае. Таким образом получается основа для логистической регрессии, где в факторы входит набор оценок, спрогнозированных методом кригинга. Кроме того, в эти факторы могут входить цветовые характеристики с аэрофотоснимка. В результате для каждой точки зоны можно вычислить вероятность превышения уровня, в случае, если она окажется близка к 1, есть основания полагать, что в такой точке величина параметра превышает пороговый уровень, а если вероятность близка к 0, есть основания считать, что значение параметра ниже порогового. Кроме того, представлен пример реализации подхода с помощью языка R на смоделированных данных. Библиогр. 8 назв. Ил. 4. Табл. 1.

Ключевые слова: экологические данные, ординарный кригинг, логистическая регрессия, язык R.

Буре Владимир Мансурович — доктор технических наук, профессор; vlb310154@gmail.com
Митрофанова Ольга Александровна — аспирант; omitrofa@gmail.com

Bure Vladimir Mansurovich — doctor of technical sciences, professor; vlb310154@gmail.com
Mitrofanova Olga Aleksandrovna — postgraduate student; e-mail: omitrofa@gmail.com

DOI: 10.21638/11701/spbu10.2016.309

V. M. Bure[1,2], *O. A. Mitrofanova*[1]

PREDICTION OF THE SPATIAL DISTRIBUTION OF ECOLOGICAL DATA USING KRIGING AND BINARY REGRESSION

[1] Agrophysical research institute, 14, Grazhdanskiy pr., St. Petersburg, 195220, Russian Federation
[2] St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

There are many ecological problems associated with the prediction of the spatial distribution of ecological parameters. The paper deals with one of these tasks. Suppose we have a set of ecological data measured by contact way (for example, plant leaf color intensity by N-tester), as well as an air photo of the object (for example, field). It is necessary to estimate the spatial distribution of ecological parameters. This paper proposes an approach to the solution of such problems with the joint use of kriging and binary regression. At first the uniform field areas (clusters) in the photo are determined using classification method. It is assumed that each selected area has a set of ecological data. Next, we will consider each zone separately. It is necessary to assess the level of the indicator in the given area. First variograms analysis is performed leading to the construction of the variogram model. Next construct a set of ecological parameter estimates is built using the method of ordinary kriging. Then, we set a threshold value of the ecological parameter for the zone under study. We introduced a variable that takes the value 1, if the parameter exceeds a threshold, and 0 otherwise. Thus we get a basis for logistic regression, where factors include a set of estimates predicted by kriging. In addition, these factors may include the color characteristics from air photos. As a result, we can calculate for each point the probability, if it will be close to 1, there is reason to believe that at this point the parameter value is greater than the threshold, and if the probability is close to 0, there is reason to assume that the parameter value is below the threshold. Furthermore, this paper provides an example of the approach for simulated data using R. Refs 8. Figs 4. Table 1.

Keywords: ecological data, ordinary kriging, logistic regression, R.

Введение. В настоящее время при исследовании многих экологических проблем важное значение имеют различные аспекты статистического анализа экологических данных, а также методы анализа цифровых изображений. Довольно часто возникают задачи, связанные с прогнозом пространственного распределения экологических параметров [1–3].

Однако не всегда требуется точная оценка значений параметра, в ряде задач достаточно оценить уровень экологических параметров в выделенной зоне поля. Изучим одну из таких задач.

Объекты и методы. Предположим, что имеется набор экологических данных $Z(x_i)$, измеренных контактным способом (агрохимия сельскохозяйственного поля, содержание азота в растениях и др.), а также аэрофотоснимок обследуемого объекта. Можно предварительно с помощью метода классификации определить однородные зоны (кластеры) поля на снимке и изучать в дальнейшем каждую зону отдельно. Необходимо оценить уровень экологического параметра в выделенной зоне поля.

Изучим два метода статистического анализа, которые использовались для решения этой задачи: ординарный кригинг и логистическую регрессию.

Кригинг. Кригинг представляет собой базовый интерполяционный геостатистический метод, который позволяет прогнозировать распределение параметра на основе набора наблюдений $Z(x_i)$. Основная формула кригинга формируется как линейная комбинация исходных данных [4]:

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i), \tag{1}$$

где $Z(x_i)$ — наблюдаемое значение в местоположении i; λ_i — неизвестный вес для него; $\hat{Z}(x_0)$ — прогнозируемое значение в местоположении x_0; n — количество наблюдений.

В ординарном кригинге предполагается, что среднее значение постоянной неизвестно. Отсутствие знания о нем накладывает на веса λ_i дополнительное требование:

$$\sum_{i=1}^{n} \lambda_i = 1. \tag{2}$$

Веса λ_i зависят от установленной модели вариограммы для установленных точек, от расстояния до местоположения прогноза и от пространственных отношений между значениями вблизи от местоположения прогноза.

Для осуществления прогноза методом кригинга необходимо провести анализ и моделирование корреляционной структуры данных (вариограммный анализ). Пространственная непрерывность данных описывается с помощью статистических моментов второго порядка. Ковариация — статистическая мера корреляции между величинами $Z(x)$ и $Z(x+h)$ в точках, разделенных вектором h:

$$C(h) = E\{[Z(x) - m(x)][Z(x+h) - m(x+h)]\}.$$

Для метода кригинга используется полувариограмма (будем называть просто вариограммой):

$$\gamma(h) = \frac{1}{2}\mathrm{Var}[Z(x) - Z(x+h)] = \frac{1}{2}E[Z(x) - Z(x+h)]^2.$$

Она связана с ковариацией следующим соотношением:

$$\gamma(h) = C(0) - C(h)$$

и характеризует пространственные отношения между наблюдениями. Чем ближе значения данных (меньше разница между ними), тем больше величина вариограммы.

Вариограмма $\gamma(h)$ оценивается на основе экспериментальной вариограммы:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2,$$

где $N(h)$ — число пар экспериментальных точек, разделенных вектором h. Важным этапом анализа является построение экспериментальной вариограммы. На рис. 1 схематично показаны основные компоненты вариограммы: наггет (самородок) c_0 — величина вариограммы при $h = 0$, порог $c + c_0$ и ранг a — предельное значение вариограммы (если оно достигается) и расстояние, на котором оно достигается, соответственно.

Для применения метода кригинга необходимо на основе экспериментальной вариограммы построить ее теоретическую модель. Приведем наиболее известные типы моделей вариограмм.

1. Сферическая модель:

$$\gamma(h) = \begin{cases} c_0 + c[\frac{3h}{2a} - \frac{1}{2}(\frac{h}{a})^3], & 0 \leqslant h \leqslant a, \\ c_0 + c, & h > a. \end{cases}$$

2. Экспоненциальная модель:

$$\gamma(h) = c_0 + c[1 - e^{-\frac{h}{a}}].$$

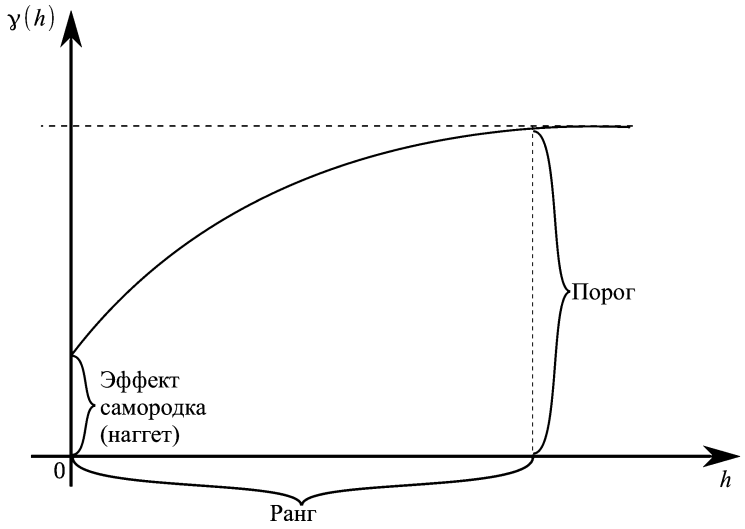

Рис. 1. Основные компоненты вариограммы

3. Гауссова модель:

$$\gamma(h) = c_0 + c[1 - e^{-\frac{h^2}{a^2}}].$$

4. Модель с эффектом дырок (скважинный эффект):

$$\gamma(h) = c_0 + c[1 - \frac{\sin(h/a)}{h/a}]. \tag{3}$$

При этом модели вариограмм могут комбинироваться.

С помощью полученных результатов вариограммного анализа необходимо найти веса оценки ординарного кригинга (1), которые минимизируют вариацию при дополнительном ограничении (2). Решение этой задачи осуществляется с использованием минимизации лагранжиана $L(x)$

$$L(x) = \sigma_z^2 + \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\tilde{\gamma}_{ij} - 2\sum_{i=1}^{N}\lambda_i\tilde{\gamma}_{i0} + 2\mu\left(\sum_{i=1}^{N}\lambda_i - 1\right).$$

Здесь $\tilde{\gamma}_{ij}$ — значение γ_{ij}, предсказанное установленной моделью вариограммы; $\tilde{\gamma}_{i0}$ — предсказанное значение между точкой i и точкой 0; μ — множитель Лагранжа; σ_z^2 — вариация ошибки Z. В результате минимизации получается линейная система из $N+1$ уравнений с $N+1$ неизвестными, которая имеет единственное решение при положительной определенности функции ковариации и отсутствии пространственно совпадающих или очень близких точек. С помощью полученных весов проводится дальнейшая оценка по формуле (1).

Логистическая регрессия. Установим пороговое значение d для рассматриваемой зоны поля и введем фиктивную переменную

$$y(x) = \begin{cases} 1, & Z(x) \geqslant d, \\ 0, & Z(x) < d. \end{cases}$$

В наблюдаемых точках экологический параметр известен, следовательно, известны и $y(x_i)$ в этих точках. Кроме того, можно спрогнозировать величины параметра в этих точках с помощью метода кригинга. Таким образом, получается основа для логистической регрессии, отражающей зависимость между вероятностью превышения порогового значения и объясняющими переменными [5, 6]:

$$P(y(x_i) = 1|\Phi_i) = p_i = \frac{1}{1 + \exp(-\Phi_i^T \beta)}. \tag{4}$$

Здесь Φ_i — факторы, объясняющие фиктивную переменную $y(x_i)$. Как один из факторов логистической регрессии предлагается ввести набор значений, предсказанных методом кригинга [7]. Кроме того, в факторы могут входить цветовые параметры со снимка, в случае, если величина экологического параметра коррелирует со значением цвета. Вектор β можно оценить методом максимального правдоподобия.

Проверку значимости построенного уравнения логистической регрессии (4) можно провести по критерию отношения правдоподобия, а также с помощью критерия Вальда [8].

Результаты и их обсуждение. Продемонстрируем применение этих методов на модельном примере, используя функции языка R. Смоделируем набор из 50 величин некоторого экологического параметра Z_i (например, показания N-тестера) на участке поля так, чтобы фиктивная переменная была следующей:

$$y(X,Y) = \begin{cases} 1; & 50 \leqslant X \leqslant 150,\ 50 \leqslant Y \leqslant 200, \\ \text{rand}(0,1); & 150 < X \leqslant 200,\ 200 < Y \leqslant 300, \\ 0; & 200 < X \leqslant 250,\ 300 < Y \leqslant 500, \end{cases}$$

где X, Y — координаты. Смоделируем также выборку объема 50 точек из нормального распределения с математическим ожиданием 0 и среднеквадратическим отклонением 1, которая будет соответствовать случайной величине ε («белый шум»), и добавим ее к смоделированному набору данных: $Z_i^* = Z_i + \varepsilon_i$. Установим порог $d = 350$. На рис. 2 схематично представлено распределение смоделированных данных на участке поля. Все расчеты проводились с использованием языка R.

Предварительно осуществлялась проверка ограничений и предположений геостатистики (стационарность и мультинормальность):

1. С помощью функции hist() строилась гистограмма частот.

2. Рассчитывались основные статистические показатели данных: среднее mean(), медиана median(), дисперсия var(), среднеквадратическое отклонение sd(), асимметрия skewness(), эксцесс kurtosis().

3. Оценивалась линейная корреляция экологического параметра с координатами с помощью функции cor.test(). Выявленные оценки указали на присутствие пространственного тренда, т. е. исходные данные противоречат гипотезе стационарности математического ожидания $m(x)$, в связи с чем появилась необходимость построения модели тренда $\hat{m}(x)$ методом множественной линейной регрессии с помощью функции lm(). После этого определили остатки (вычли модельные значения из исходных данных), все дальнейшие методы анализа применялись к остаткам.

4. Осуществлялась проверка мультинормальности, строился q–q график с помощью функций *qq*norm() и *qq*line(), проверялась гипотеза одномерной нормальности благодаря критерию Колмогорова–Смирнова ks.test(). Полученные результаты не противоречат гипотезе многомерной нормальности.

Рис. 2. Базовая карта наблюдений

После предварительного анализа осуществлялся вариограммный анализ:

1. Строилась гистограмма расстояний, оценивался разброс расстояний.

2. Строились h-графики с помощью функции hscat() с направлениями 0, 45, 90 и 135°.

3. Строилась поверхность экспериментальной вариограммы с помощью функций variogram() и plot(). Поверхность вариограммы после сглаживания иллюстрирует рис. 3, анизотропия не наблюдается.

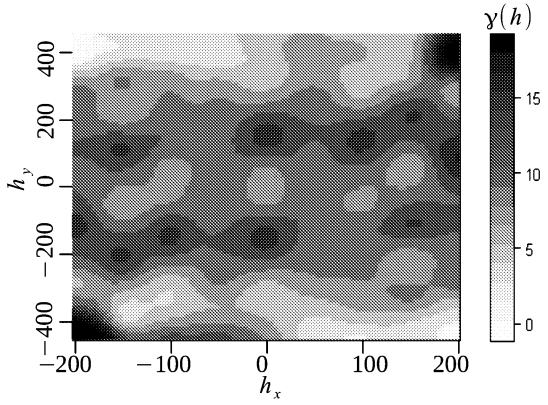

Рис. 3. Поверхность вариограммы

4. Вариограмма оценивалась по четырем направлениям: 0, 45, 90 и 135°. На основе этой оценки устанавливалась модель вариограммы с помощью функции vgm(), использовалась модель с эффектом дырок (3). Экспериментальную вариограмму по четырем направлениям с установленной моделью вариограммы иллюстрирует рис. 4.

На основе результатов вариограммного анализа применялся ординарный кригинг (1). Поочередно из набора смоделированных наблюдений исключалось одно значение, после чего оно оценивалось методом кригинга с помощью функции krige(). Таким

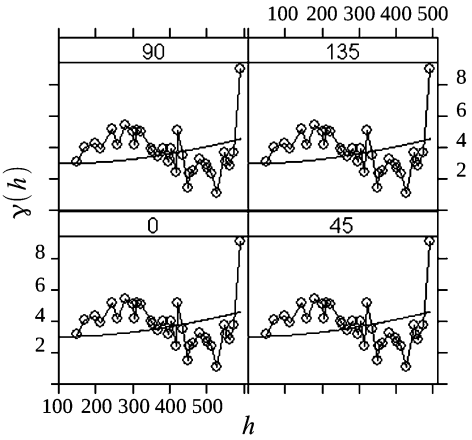

Рис. 4. Вариограмма по четырем направлениям

образом, был выявлен набор значений экологического параметра, предсказанных методом ординарного кригинга в наблюдаемых точках.

С помощью выявленных данных строились три логит-модели на основе функции glm(). Оценка значимости моделей проводилась по критерию отношения правдоподобия. Исходные данные для построения следующие: зависимая переменная $T = 1$, если показание N-тестера более или равно 350, иначе принимает значение 0; в качестве объясняющих переменных рассматривались переменные X и Y — координаты наблюдений, а также N_{pred} — предсказанные методом кригинга величины параметра в наблюдаемых точках.

***Модель 1*:** зависимая переменная T, объясняющие переменные X и Y. В результате построения получили уравнение

$$P(T = 1) = \frac{1}{1 + \exp(-8.754 + 0.0314X + 0.0193Y)}.$$

При этом коэффициент хи-квадрат равен 37.36926, уровень значимости составляет 3.840071e-09, уровни значимости смоделированных коэффициентов — меньше 0.05. Таким образом, построенная модель статистически значима.

***Модель 2*:** зависимая переменная T, объясняющие переменные X, Y и N_{pred}. В результате построения получили уравнение

$$P(T = 1) = \frac{1}{1 + \exp(-11554.652 + 0.2814X + 0.2862Y + 32.7542N_{\text{pred}})}.$$

При этом коэффициент хи-квадрат равен 51.0379, уровень значимости составляет 2.355682e-11, уровни значимости смоделированных коэффициентов — меньше 0.05. Таким образом, построенная модель статистически значима.

***Модель 3*:** зависимая переменная T, объясняющая переменная N_{pred}. В результате построения получили уравнение

$$P(T = 1) = \frac{1}{1 + \exp(738.3974 - 2.112N_{\text{pred}})}.$$

При этом коэффициент хи-квадрат равен 30.26213, уровень значимости составляет 1.945903e-08, уровни значимости смоделированных коэффициентов — меньше 0.05. Таким образом, модель 3 статистически значима.

На заключительном этапе сравнивались данные модели. Так как все три модели вложенные, предварительное сравнение моделей осуществлялось с помощью функции anova(). В результате пришли к выводу, что полная модель работает лучше сокращенных. Кроме того, была создана дополнительная тестовая база из 50 точек: поочередно из набора наблюдений исключалась одна точка и осуществлялся анализ, как предскажут в этой точке значение вероятности $P(T = 1)$ все три логит-модели. Как вытекает из таблицы, где представлена выборка из 10 точек построенной тестовой базы, в 37 точках вторая модель показала себя лучше сокращенных моделей.

Выборка из тестовой базы логит моделей

№	X	Y	Z	T	N_{pred}	Модель 1	Модель 2	Модель 3
1	100	50	352	1	351.1794	0.9902502	**0.9999325**	0.9623315
2	250	50	348	0	350.2188	0.6091639	**0.3478838**	0.827778
3	200	100	347	0	350.2588	0.7108186	**0.07790791**	0.8410091
4	150	150	354	1	349.9958	0.7376028	**0.9966405**	0.6703221
5	150	200	353	1	349.6525	0.5135569	**0.9235411**	0.4926213
6	50	300	353	1	349.5278	0.7721649	**0.9980404**	0.4270269
7	150	300	348	0	348.954	0.1547047	**0.04604387**	0.2027392
8	50	400	345	0	349.0604	0.4353104	**0.00086428**	0.2452617
9	100	450	347	0	348.1283	0.0458002	**0.00742519**	0.0420072
10	150	500	348	0	347.3104	0.0036139	**0.00145936**	0.0075487

Заключение. В работе предложен подход к решению актуальных экологических проблем, связанных с прогнозом пространственного распределения экологических параметров, а также продемонстрирован пример реализации этого подхода на смоделированных данных средствами языка программирования R. На основе проведенного исследования можно прийти к выводу, что описанный подход:

1) доступный, недорогой и достаточно точный прием прогноза пространственного распределения экологического показателя;

2) дает возможность спрогнозировать значение экологического параметра в каждой точке поля;

3) дает возможность решать ряд актуальных экологических проблем: мониторинг состояния растений, прогноз урожайности, дифференцированное внесение азотных удобрений и др.

На основе полученных в эксперименте результатов можно полагать, что целесообразно использовать полную логит-модель, однако этот подход требуется рассмотреть на ряде дополнительных примеров.

Литература

1. *Буре В. М.* Методологические аспекты статистического анализа в точном земледелии // Докл. Рос. академии сельскохоз. наук. 2007. № 6. С. 54–56.

2. *Митрофанова О. А., Буре В. М., Канаш Е. В.* Математический модуль для автоматизации колориметрического метода оценки обеспеченности растений азотом // Вестн. С.-Петерб. ун-та. Сер. 10. Прикладная математика. Информатика. Процессы управления. 2016. Вып. 1. С. 85–91.

3. *Якушев В. П., Буре В. М.* Оценка биоэквивалентности двух участков на сельскохозяйственном поле // Докл. Рос. академии сельскохоз. наук. 2006. №. 5. С. 38–40.

4. *Демьянов В. В., Савельева Е. А.* Геостатистика: теория и практика. М.: Ин-т проблем безопасности развития атомной энергетики РАН; Наука, 2010. 327 с.

5. *Буре В. М.* Методология применения бинарной регрессии в точном земледелии // Математические модели в теоретической экологии и земледелии: материалы Междунар. семинара, посвященного памяти профессора Ратмира Александровича Полуэктова (Полуэктовские чтения). 2014. С. 118–121.

6. *Якушев В. П., Буре В. М., Парилина Е. М.* Бинарная регрессия и ее применение в агрофизике. СПб.: Агрофиз. ин-т, 2015. 36 с.

7. *Fernandes G. B., Artes R.* Spatial dependence in credit risk and its improvement in credit scoring // European J. of Operational Research. 2016. N 249. P. 517–524.

8. *Буре В. М., Парилина Е. М.* Теория вероятностей и математическая статистика. СПб.: Изд-во «Лань», 2013. 416 с.

Для цитирования: Буре В. М., Митрофанова О. А. Прогноз пространственного распределения экологических данных с применением кригинга и бинарной регрессии // Вестник Санкт-Петербургского университета. Сер. 10. Прикладная математика. Информатика. Процессы управления. 2016. Вып. 3. С. 97–105. DOI: 10.21638/11701/spbu10.2016.309

References

1. Bure V. M. Metodologicheskie aspekty statisticheskogo analiza v tochnom zemledelii [Methodological aspects of statistical analysis in precision agriculture]. *Russian Agricultural Sciences,* 2007, no. 6, pp. 54–56. (In Russian)

2. Mitrofanova O. A., Bure V. M., Kanash E. V. Matematicheskii modul' dlia avtomatizatsii kolorimetricheskogo metoda otsenki obespechennosti rastenii azotom [Mathematical module to automate the colorimetric method for estimating nitrogen status of plants]. *Vestnik of Saint Petersburg University. Series 10. Applied mathematics. Computer science. Control processes*, 2016, issue 1, pp. 85–91. (In Russian)

3. Yakushev V. P., Bure V. M. Otsenka bioekvivalentnosti dvukh uchastkov na sel'skokhoziaistvennom pole [Approach to evaluating bioequivalence of two plots on an agricultural field]. *Russian Agricultural Sciences,* 2006, no. 5, pp. 38–40. (In Russian)

4. Dem'ianov V. V., Savel'eva E. A. *Geostatistika: teoriia i praktika* [*Geostatistics: theory and practice*]. Moscow, Nuclear Safety Institute of the Russian Academy of Sciences, Nauka Publ., 2010, 327 p. (In Russian)

5. Bure V. M. Metodologiia primeneniia binarnoi regressii v tochnom zemledelii [Methodology of using binary regression in precision agriculture]. *Poluektov's readings*, 2014, pp. 118–121. (In Russian)

6. Yakushev V. P., Bure V. M., Parilina E. M. *Binarnaia regressiia i ee primenenie v agrofizike* [*Binary regression and its application in agrophysics*]. Saint Petersburg, Agrophys. Institute Publ., 2015, 36 p. (In Russian)

7. Fernandes G. B., Artes R. Spatial dependence in credit risk and its improvement in credit scoring. *European J. of Operational Research,* 2016, no. 249, pp. 517–524.

8. Bure V. M., Parilina E. M. *Teoriia veroiatnostei i matematicheskaia statistika* [*Probability theory and mathematical statistics*]. Saint Petersburg, Lan' Publ., 2013, 416 p. (In Russian)

For citation: Bure V. M., Mitrofanova O. A. Prediction of the spatial distribution of ecological data using kriging and binary regression. *Vestnik of Saint Petersburg University. Series 10. Applied mathematics. Computer science. Control processes*, 2016, issue 3, pp. 97–105. DOI: 10.21638/11701/spbu10.2016.309

Статья рекомендована к печати проф. Л. А. Петросяном.

Статья поступила в редакцию 10 апреля 2016 г.

Статья принята к печати 26 мая 2016 г.